OctoPack: Instruction Tuning Code Large Language Models

Year: 2023 | Citations: 176 | Authors: Niklas Muennighoff, Qian Liu, Qi Liu

---

Abstract

Finetuning large language models (LLMs) on instructions leads to vast performance improvements on natural language tasks. We apply instruction tuning using code, leveraging the natural structure of Git commits, which pair code changes with human instructions. We compile CommitPack: 4 terabytes of Git commits across 350 programming languages. We benchmark CommitPack against other natural and synthetic code instructions (xP3x, Self-Instruct, OASST) on the 16B parameter StarCoder model, and achieve state-of-the-art performance among models not trained on OpenAI outputs, on the HumanEval Python benchmark (46.2% pass@1). We further introduce HumanEvalPack, expanding the HumanEval benchmark to a total of 3 coding tasks (Code Repair, Code Explanation, Code Synthesis) across 6 languages (Python, JavaScript, Java, Go, C++, Rust). Our models, OctoCoder and OctoGeeX, achieve the best performance across HumanEvalPack among all permissive models, demonstrating CommitPack's benefits in generalizing to a wider set of languages and natural coding tasks. Code, models and data are freely available at https://github.com/bigcode-project/octopack.